# 应用于问答系统的 Lucene 相似度检索算法改进

白 菊[1],何聚厚[2]
(1. 现代教学技术教育部重点实验室,陕西 西安 710062;
2. 陕西师范大学 计算机科学学院,陕西 西安 710119)

**摘 要**:Lucene 在文本检索和搜索领域有着广泛的应用,相似度评分算法是其搜索引擎的核心部分之一。而在问答系统中,也要用到检索功能,相似度评分算法也是其核心部分之一。那么能否对 Lucene 的相似度评分算法进行改进,使其在问答系统的领域也能得到很好的应用。针对上述提出的问题,结合问答系统中问句简短、包含信息量少的特点,引入外部词典对查找的关键词进行扩展,分析检索词项的语义相似度以及将词项位置关系的特征应用到 Lucene 中。在 Lucene 的基础上,对其语义相似度算法进行改进,提出了一种新的语义相似度评分算法。该算法考虑了词项位置关系和语义理解,能够更好地应用于问答系统。实验结果表面,提出的相似度算法能有效地提高自动问答系统的回答准确率。

**关键词**:Lucene;相似度;问答系统;语义
**中图分类号**:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)11-0079-04
doi:10.3969/j.issn.1673-629X.2017.11.017

# Improvement of Lucene Similarity Search Algorithm Applied in Question Answering System

BAI Ju[1],HE Ju-hou[2]
(1. Key Laboratory of Modern Teaching Technology of Ministry of Education,Shaanxi Normal University,
Xi'an 710062,China;
2. School of Computer Science,Shaanxi Normal University,Xi'an 710119,China)

**Abstract**:Lucene has a wide range of applications in the field of text retrieval and search,and the similarity score algorithm is one of the key parts of its search engine. And in the question answering system,the search function is also used,and the similarity score algorithm is also one of the key parts of its search engine. It is possible to improve the similarity score algorithm of the Lucene so that it can be widely used in the field of question answering system. In view of this problem,combined with the question answering system in the characteristic of brief question and small amount of information,the external dictionary is introduced to expand the searched key words,analysis and retrieval of semantic similarity of words,application of lexical position relationship feature in Lucene. On the basis of Lucene,its semantic similarity algorithm is improved,and a new one is proposed which can be better applied in question answering system in consideration of lexical position relationship and semantic understanding. Experimental results show that the proposed algorithm can effectively improve the accuracy of the question answering system.

**Key words**:Lucene;similarity;question answering system;semantics

## 0 引 言

Lucene 是用 Java 语言实现的开放源代码的全文检索引擎工具包,是 Apache 软件基金会 Jakarta 项目组下的一个子项目。Lucene 以其索引结构优异、开源特性、高性能、易使用等特点,广泛应用于 Web、文本检索等领域,以及各种软件系统中,如开源软件 Eclipse 的搜索功能等[1]。虽然 Lucene 有着广泛的应用,但也存在不足之处。例如,Lucene 内部默认的是基于词频的分析检索函数来考察检索文本之间的相似性[2],很少有考虑词项语义的相似度,也没有考虑到词语位置之间的关系对搜索准确的影响。而且 Lucene 是一个开源的检索框架,并不是一个完整的搜索引擎,它只是一个工具包[3]。因此,如果能对 Lucene 的检索函数加以改进,并结合问答领域的特点,则其在问答领

**收稿日期**:2016-11-16 **修回日期**:2017-03-30 **网络出版时间**:2017-08-01
**基金项目**:教育部-中国移动科研基金项目(MCM20150604)
**作者简介**:白 菊(1990-),女,硕士研究生,研究方向为知识工程与智能教学系统;何聚厚,博士,副教授,研究方向为知识工程与智能系统。

域也能有很好的应用。

根据以上问题,将检索词项语义的相关应用考虑到该检索函数中。该函数改进了 Lucene 忽略语义信息而只考虑基于词频的检索方法所引起的检索不精确的问题[4]。在将词扩展应用到该函数中的同时,也将词项位置关系特征考虑到该检索函数中。这样不但实现了 Lucene 的相似度算法的改进,也使其能更好地适应问答领域中问句短、信息量少的特点。

## 1 Lucene 的相似度评分算法分析

Lucene 的核心部分之一就是相似度评分算法,是用来衡量查询串和检索文档相似程度的一种算法。Lucene 使用一种基于向量空间模型(SVN)的 TF-IDF 方法来计算文档的相似度。TF-IDF 方法基于词频分析来考虑检索文档的相似度,它综合考虑的是这个词对不同文档的分辨能力和不同的词在所有文档中的出现频率[5]。

Lucene 内部的相似度评分算法如式(1)所示:

$$\text{Score}(d) = \sum_{t \text{ in } q} \text{tf}(t \text{ in } d) \times \text{idf}(t)^2 \times \text{boost}(t.\text{field in } d) \times \text{lengthNorm}(t.\text{field in } d) \times \text{coord}(q,d) \times \text{queryNorm}(q) \quad (1)$$

其中, $\text{idf}(t)^2$ 表示根据词项 $t$ 在倒排索引中出现的频率; $\text{tf}(t \text{ in } d)$ 表示文档 $d$ 中词项 $t$ 出现的频率; $\text{boost}(t.\text{field in } d)$ 表示词项 $t$ 所在的域的加权因子; $\text{lengthNorm}(t.\text{field in } d)$ 是在索引过程中计算出来并存储在索引中的,表示域的标准化值,表示在某个域中词项的个数; $\text{coord}(q,d)$ 是一个协调因子,其取值大小由问答中包含的查询词项的多少决定。包含的查询词项越多,该值就越大; $\text{queryNorm}(q)$ 表示每个查询词项的标准值,即每个查询词项权重的平方和。

从式(1)可以看出 Lucene 内部的相似度评分算法的特点:

(1)一个文档中包含该查询词项的频率越高,该文档的得分就越高;

(2)查询词项在文档中的位置并不重要;

(3)在一个命中文档中,如果除了该查询词之外,其他的词越多,该文档得分越少[6-7]。

但在多数情况下,文档与词项的相似程度不但与词项出现的频率有关,还与词项的位置关系特征以及词义有关。例如,不同文档中有下面两句话:

S:mooc 发展的主要问题是市场环境和体制问题。

R:这种病的主要根源是生活的环境问题造成的。

对于查询“环境问题”,在文档 R 中的查找是完全匹配的,因此,R 文档的得分应该比 S 文档高,也更符合查询者的要求。但是由于这两句话中包含的词项“问题”在文档 R 中只出现一次,而在文档 S 中出现两次,所以由 Lucene 检索得出的结果反而是 S 文档的得分高于 R 文档的得分[8-9]。

## 2 算法的改进

Lucene 内部缺省实现的相似度检索函数不考虑词项的含义,也不考虑词项出现的顺序,而是将文本看作一个容纳词项的袋子。文本特征向量由文本中出现的词项在文本中的频率以及该词项在整个文本集中出现的频率表示。每一篇文本建模为由文中出现的 $n$ 个加权词项组成的向量。该方法基于以下两点[10]:

(1)词频(Term Frequency):某个词项在文本中出现的频率越高,则它和该文本的相关度越高;但在很多特定的语言环境下,有许多特定的词不具备这种特性,从而应将其排除,如英文的“she”和“he”,中文的“的”和“得”。

(2)逆文本词项频率(Inverse Document Frequency):某个词项在文本集合的多篇文本中出现的频率越高,则该词项的区分度越差。例如,在包含 1 500 个文本集的集合中,某个词项 S 在 300 篇文章中都有出现,而另一个词项 R 只在 30 篇文章中出现,则词项 R 比 S 有更好的区分度。通过对文本集中的每一个词项都进行上述分析,然后得到每一篇文章中每一个词项的 TF-IDF 值[11]。再利用这些 TF-IDF 值为每一篇文章建立一个空间向量模型,通过计算 Jaccard 系数或向量间的余弦相似度来表示检索与文本之间的相似性。最终根据检索文档与用户查询之间的相似度值的高低排序,将检索结果返回给用户[12]。

### 2.1 语义改进

尽管上述 Lucene 内部相似度评分算法在实践应用中效果较好,但未能捕捉到文本的语义信息。而在自动应答系统中,用户提出的问句本来就比较短,能捕捉到的信息也比较少,如果不考虑语义信息,则给用户返回的回答的准确率可想而知[13]。例如,用户提问有关电脑的问题,而电脑也称计算机,如果仅用 Lucene 中只考虑词频而不考虑语义的方法,只能搜到有“电脑”这个词的回答,而在只有“计算机”这个词的答案是找不到的[14]。如果考虑检索词项的语义信息,则能更准确地获取用户的检索信息[15]。

对以上提出的问题,对式(1)的相似度算法进行改进,改进后的相似度算法函数如下所示:

$$\text{SimScore}(d) = \text{Sim}\left[\sum_{t \text{ in } q} \text{tf}(t \text{ in } d)\right] \times \text{idf}(t)^2 \times \text{boost}(t.\text{field in } d) \times \text{lengthNorm}(t.\text{field in } d) \times \text{coord}(q,d) \times \text{queryNorm}(q) \quad (2)$$

其中, Sim[ tf($t$ in $d$) ] 表示在进行查询之前,先对词项 $t$ 进行扩展,将与词项 $t$ 相似的词项加到查询词项中之后再进行查询。在对词项 $t$ 进行扩展时,引入外部词典 WordNet 对词项进行相似度查询。

### 2.2　词项位置改进

词项的位置关系在问答系统中也占有非常重要的位置,对于本来信息量就比较少的问句,词项的位置关系特征对回答准确率的影响可能就至关重要。而词项的位置关系特征不仅与词项出现的频率有关,还与词语位置的关系特征有关[15]。文中将词项位置的关系特征分为三种:当距离为 1 时,这两个词是直接相邻的;当距离大于 0 小于 1 时,该词是去掉停用词后相邻的;当距离等于 0 时,这两个词是不相邻的。因此,词项的位置关系特征可进一步表示为词项间的距离关系。为了更好地体现这一关系,引入“词项位置相邻相似度”来反映查询词项与检索文档中的词项在相邻性关系上的相似程度。

在对 Lucene 的评分机制函数进行改进前,首先对分词处理后此项之间的相邻程度进行标注。若两个词在分词前后都是相邻的(即中间没有去掉的字或词),则两个词之间的距离等于 1;如果在分词后中间有停用词或字,去掉停用词或字后是相邻的,则两词之间的距离等于 0.7;否则两词之间的距离等于 0。词项距离得分如下所示:

$$\mathrm{OrdScore}=\begin{cases}1 & \text{没有停用词相邻}\\ 0.7 & \text{去停用词后相邻}\\ 0 & \text{不相邻}\end{cases} \tag{3}$$

根据以上分析,对 Lucene 相似度评分函数进行改进,如下所示:

$$\mathrm{NewScore}(d)=\alpha\times\mathrm{SimScore}(d)+\beta\times\mathrm{OrdScore}(d) \tag{4}$$

其中, $\mathrm{SimScore}(d)$ 是加入外部字典后计算的相似度得分; $\mathrm{OrdScore}(d)$ 是考虑了词位置后计算的相似度得分; $\alpha+\beta=1$,经过实验,当 $\alpha=0.6$, $\beta=0.4$ 时,搜索结果最为有效。

## 3　实验结果与分析

### 3.1　引入外部词典

(1)在数据库里面同时有“mooc”和“慕课”两个词语,但没有引入同义词词典时的搜索结果如图 1 所示。

从图 1 可以看出,没有加查询扩展之前,电脑无法识别出“慕课”和“mooc”是同一个意思,所以在查找时输入“mooc”就不会出现“慕课”的相关回答。

(2)引入外部词典,进行查询扩展之后的搜索结果如图 2 所示。

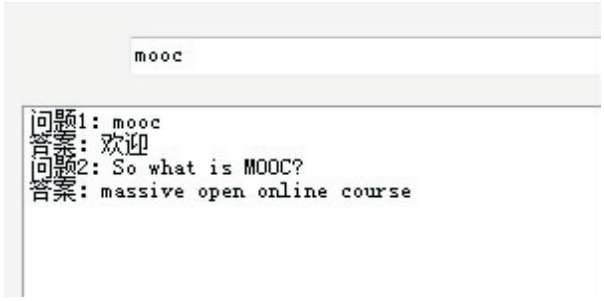


图 1　未引入同义词字典的搜索结果

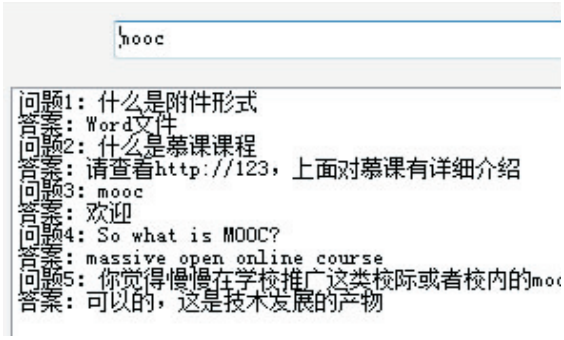


图 2　引入外部词典的搜索结果

从图 2 可以看出,在引入外部词典进行查询扩展之后,输入“mooc”后,和“慕课”相关的答案也会得出。这样的查询结果更符合用户的需求。

### 3.2　考虑词位置关系

(1)没有考虑词位置关系前的搜索结果如图 3 所示。

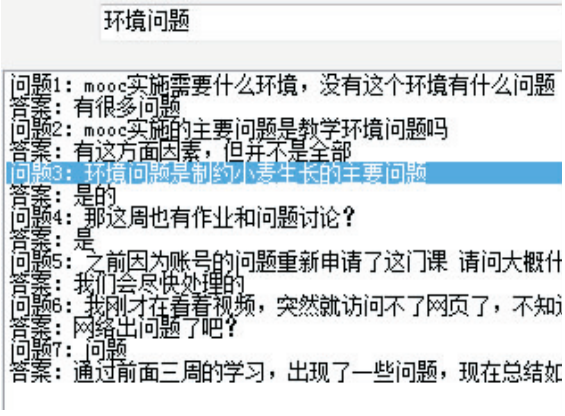


图 3　未考虑词位置关系的搜索结果

(2)考虑词位置后的搜索结果如图 4 所示。

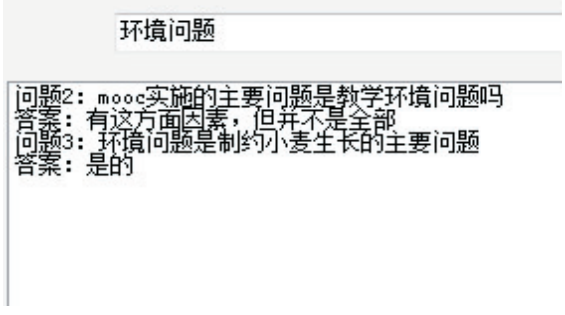


图 4　考虑词位置关系的搜索结果

从上面的搜索结果可以看出,在没有考虑词位置前,“环境问题”被分成“环境”和“问题”两个没有任何位置关系的词分别进行查找,查找结果只要有“环境”或“问题”的相关答案都会出来。加上词位置关系后,当查找到“环境”和“问题”两个词语位置相邻的相关答案时,得分会更高,会显示在更前面,这样查找的准确率明显高于前面查找的准确率,更符合用户的查找要求。

### 3.3 综合结果比较

任选七个问句,出现相关的前 3 个答案时所包含的答案条数如表 1 所示。

表 1 算法比较

| 算法 | Mooc 是什么 | 这堂课上多长时间 | 这门课成绩怎么算 | 学习网址是什么 | 视频打不开怎么办 | 这门课程只在网上学习吗 | 平时作业没做怎么办 |
|---|---|---|---|---|---|---|---|
| 改进前相似度算法 | 12 | 9 | 10 | 7 | 4 | 11 | 14 |
| 改进后相似度算法 | 4 | 9 | 7 | 7 | 3 | 5 | 7 |

折线图如图 5 所示。

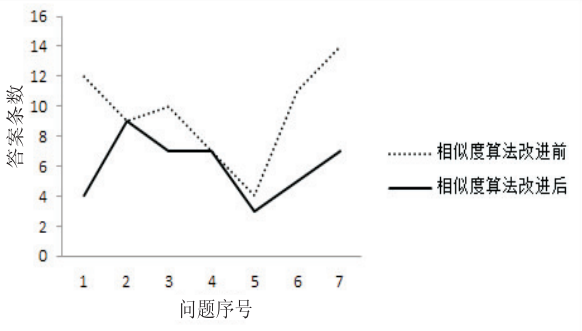


图 5 相似度改进前后结果对比

从图 5 中可以看出,在对 Lucene 的相似度算法改进后,搜索的答案明显优于未改进前。

## 4 结束语

在自动应答系统领域,语义相似度计算是一个极其重要的问题。文中对 Lucene 内部的相似度评分算法进行了阐述,并在此基础上对其进行改进。引入词项位置关系特征和语义相似度信息来提高检索的准确率。检索结果表明,提出的方法可行、有效。虽然该算法有了一定的改进,但依然存在不足;在该算法中,引入的外部词典都需要提前将近义词等组织好,比较麻烦。需要寻找更简单有效的方法,使自动应答系统能自动识别近义词,而无需外部引入,这将是下一阶段研究要考虑的问题。

**参考文献:**

[1] 李永春,丁华福. Lucene 的全文检索的研究与应用[J]. 计算机技术与发展,2010,20(2):12-15.
[2] 吴代文,杨方琦. Lucene 在数据库全文检索中的性能研究[J]. 微计算机应用,2011,32(6):53-59.
[3] 张 俊,李鲁群,周 熔. 基于 Lucene 的搜索引擎的研究与应用[J]. 计算机技术与发展,2013,23(6):230-232.
[4] 杨 彬. 基于 Lucene. NET 的局域网全文搜索引擎的设计与实现[D]. 四川:电子科技大学,2014.
[5] Pirro G,Talia D. An approach to ontology mapping based on the Lucene search engine library[C]//18th international workshop on database and expert systems applications. [s. l.]:IEEE,2007:407-411.
[6] 余正涛,樊孝忠,宋丽哲. 基于问句语料库的受限领域自动应答系统[J]. 计算机工程与应用,2003,39(36):28-30.
[7] 张 宏. 基于本体的农业自动应答系统关键技术研究[D]. 保定:河北农业大学,2007.
[8] 王泽贤. 基于 Lucene 的书目搜索相似度评分算法改进研究[J]. 图书情报工作,2014,58(4):94-98.
[9] 丁兆贵,金 敏. 基于 Lucene 的个性化搜索引擎研究与实现[J]. 计算机技术与发展,2011,21(2):105-108.
[10] 袁亚静. 基于查询扩展的微博客服自动应答系统[D]. 北京:北京邮电大学,2015.
[11] 索红光,孙 鑫. 针对中文检索的 Lucene 改进策略[J]. 计算机应用与软件,2009,26(6):175-177.
[12] 任树怀. LUCENE 搜索算法剖析及优化研究[J]. 图书馆杂志,2014,33(12):17-23.
[13] 王 欢,孙瑞志. 基于领域本体和 Lucene 的语义检索系统研究[J]. 计算机应用,2010,30(6):1655-1657.
[14] 白培发,王成良,徐 玲. 一种融合词语位置特征的 Lucene 相似度评分算法[J]. 计算机工程与与应用,2014,50(2):129-132.
[15] 宋 佳,诸云强,刘润达. 一种基于 Lucene 改进的全文检索工具包[J]. 计算机工程与应用,2008,44(4):172-175.